JURNAL INFOTEL
Informatika - Telekomunikasi - Elektronika
Website Jurnal : http://ejournal.st3telkom.ac.id/index.php/infotel
ISSN : 2085-3688; e-ISSN : 2460-0997

Perhitungan Kemiripan Term *Co-occurrence* Berdasarkan *Cluster* Dokumen untuk Pengembangan Thesaurus Bahasa Arab

Dika Rizky Yunianto[1], Agus Zainal Arifin[2]
[1,2]Jurusan Teknik Informatika, Fakultas Teknologi Informasi, Institut Teknologi Sepuluh Nopember
[1,2]Jalan Raya ITS 60111, Surabaya, Jawa Timur, Indonesia
Email korespondensi: dika15@mhs.if.its.ac.id

Dikirim 17 Januari 2017, Direvisi 29 Januari 2017, Diterima 1 Februari 2017

Abstrak – Salah satu cara dalam pembentukan thesaurus adalah dengan cara menghitung nilai kemiripan *term*. Untuk mendapatkan nilai kemiripan tersebut dapat dilakukan dengan pendekatan *co-occurrence*, yaitu melihat frekuensi kemunculan bersama *term-term* tersebut. Frekuensi tersebut dilihat dari seberapa banyak *term* tersebut muncul bersama pada dokumen-dokumen corpus. Setiap dokumen-dokumen yang terdapat pada corpus memiliki konten atau topik yang berbeda-beda sehingga *term-term* yang berada pada dokumen suatu topik akan memiliki konteks yang berbeda dengan *term-term* pada dokumen dengan topik lainnya. Oleh sebab itu, paper ini mengusulkan metode baru dalam perhitungan kemiripan *term* dengan *co-occurrence* yang memperhatikan klaster dari dokumen pada pengembangan thesaurus Bahasa Arab. Dokumen-dokumen *corpus* akan di-*clustering* untuk mengelompokkan berdasarkan kedekatan konten dari dokumen tersebut. Untuk mendapatkan nilai kemiripan *term* dilakukan perhitungan *clusterweight* dengan memanfaatkan nilai dari *inverse class frequency* setiap *term* terhadap klaster yang ada. Thesaurus dibentuk dengan melihat nilai hasil perhitungan kemiripan *term* tersebut. Thesaurus yang dibentuk dengan metode usulan berhasil meningkatkan relevansi antar *term* yang dibuktikan dengan hasil percobaan memiliki nilai precision tertinggi sebesar 63,3%, recall sebesar 78,6%, dan f-measure sebesar 50%.

Kata kunci – Thesaurus, *Co-occurrence*, Kemiripan *Term*, Bahasa Arab, Klaster Dokumen

Abstract - Generating automatically is by calculating the similarity value term. To get the value of the similarity can be carried out with the co-occurrence approach is to see the frequency of occurrence along these terms. The frequency of how much these terms occurrence on documents corpus. Each of the documents contained in the corpus have content or topics vary. So, the terms that are in the document a specific topic will have a different context with the terms of the document with other topics. Therefore, this paper proposes a new method of measurement term similarity with co-occurrence based on cluster of documents on the generate of Arabic thesaurus. The documents will be in the corpus clustering to group by the proximity of the content of the document. To get the term similarity value calculation clusterweight by leveraging the value of inverse class frequency of each term to an existing cluster. Thesaurus is formed by looking at the value of the calculation result of the similarity term. Thesaurus formed by the proposed method succeeded in improving inter-term relevance is evidenced by the experimental results have a precision value of 63,3%, amounting to 78,6% recall and F-measure by 50%.

Keywords - Thesaurus, Co-occurrence; Term Similarity; Arabic; Document Cluster

I. PENDAHULUAN

Pada hakekatnya, temu kembali informasi harus dapat menampilkan dokumen-dokumen yang relevan sesuai dengan keiinginan pengguna pada proses pencarian dokumen. Terdapat permasalahan di mana kata kunci atau *query* yang digunakan untuk melakukan pencarian dokumen memiliki makna yang berbeda-beda melihat batas keamampuan pengguna dalam pemilihan kata-kata yang digunakan dalam pencarian [1]. Perbedaan makna sebuah kata atau *term*

merupakan permasalahan yang terus dikaji dalam bidang temu kembali informasi agar dokumen-dokumen yang dihasilkan dalam pencarian dokumen relevan dengan keiinginan pengguna. Sistem harus dapat mengatasi permasalahan ambiguitas suatu kata serta harus dapat mengatasi ketidak cocokan antara dokumen dengan *query* dari pengguna. Untuk mengatasi hal tersebut terdapat *tools* atau kamus yang dinamakan dengan thesaurus [2].

Thesaurus merupakan kamus yang minimal berisi daftar kemiripan *term* dan dapat digunakan sebagai alat dalam melakukan *query expansion* sehingga meningkatkan relevansi hasil pencarian [3]. Yang dimaksud dengan kemiripan *term* bukan hanya *term* yang memiliki artian sama saja, namun *term-term* yang memiliki hubungan semantik atau kemiripan berdasarkan konsep konteks atau representasi objek yang sama [4]. Sebagai contoh lain penggunaan thesaurus juga dapat dikembangkan sebagai *tools* untuk melakukan ekspansi dalam melakukan cek *plagiarism* suatu dokumen. Di mana penelitian sebelumnya telah melakukan cek *plagiarism* dengan menghimpun kata-kata yang tersebar pada dua dokumen [5].

Terdapat dua cara dalam melakukan pembangunan kamus thesaurus yaitu, pembangunan secara manual dan secara otomatis. Pembangunan thesaurus secara manual memiliki permasalahan pada lama waktu proses pembangunan serta sumber daya yang dibutuhkan. Oleh sebab itu dibutuhkan pembangunan thesaurus dengan cara otomatis untuk menekan biaya dan efektifitas waktu [1]. Pembangunan thesaurus secara otomatis memiliki banyak cara, salah satunya dengan dengan menghitung kemiripan *term* secara statistikal. *Pointwise mutual information* dan *dice* yang dapat digunakan untuk menghitung kemiripan *term* secara simetris [6].

Penelitian yang dilakukan oleh [3] untuk membangun thesaurus Bahasa Arab secara otomatis dengan pendekatan statistikal *co-occurrence*. Teknik *co-occurrence* digunakan untuk menemukan kemiripan antar *term* dalam melakukan pembangunan kamus thesaurus. Bahasa Arab digunakan sebagai studi kasus dalam pembangunan kamus thesaurus tersebut dikarenakan Bahasa Arab digunakan pada 23 negara bahasa resminya, dan hampir 422 juta orang menggunakan Bahasa Arab sebagai Bahasa keseharian. Selain itu Bahasa Arab memiliki morfolologi yang kompleks sehingga masih sedikit pengembangannya. [1]. Teknik *co-occurrence* yang digunakan untuk membangun thesaurus Bahasa Arab merupakan perhitungan kemiripan antar *term* secara asimetris dengan melihat kemunculan bersama kedua term [7].

Beberapa peneliti mengatakan bahwa perhitungan kemiripan secara asimetris lebih baik dibandingkan perhitungan secara simetris. Hal tersebut dikarenakan perhitungan secara simetris akan menyebabkan keberulangan atau kemunculan *term-term* yang bersama akan terhitung lebih sering, sehingga tidak begitu membantu dalam melakukan eksplorasi kata kunci atau *query expansion* pada saat pencarian dokumen [3].

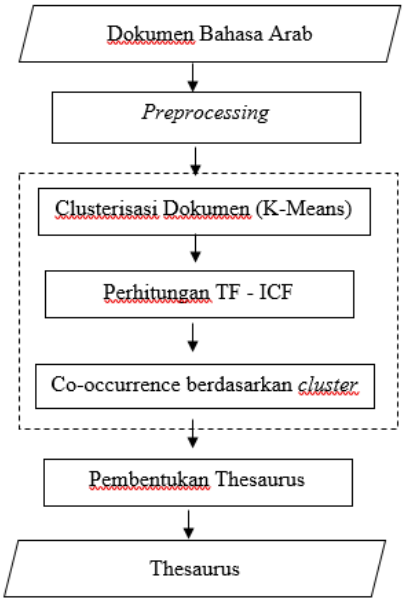
Teknik *co-occurence* dilakukan dengan memperhatikan frekuensi kemunculan bersama kedua *term*. Frekuensi kemunculan tersebut dihitung berdasarkan dokumen-dokumen yang ada pada *corpus*. Namun, setiap dokumen-dokumen pada corpus memiliki bahasan atau topiknya sendiri-sendiri. Hal tersebut mengakibatkan perbedaan konteks terhadap term-term yang berada pada topik yang berbeda. Sebagai contoh kata "صلي" pada topik "sholat" memiliki artian proses ibadah yang dinamakan sholat. Namun kata tersebut akan berbeda makna pada topik "nikah" yang memiliki artian berdo'a. Dari hal tersebut dapat disimpulkan perbedaan topik akan mengakibatkan perbedaan makna *term-term* pada topik tersebut.

Hubungan term terhadap topik dokumen atau cluster dokumen dapat dilihat dari nilai keinformatifan *term* tersebut. Nilai keinformatifan suatu *term* tidak hanya dapat dilihat dari sisi dokumen saja, melainkan juga dapat dilihat dari klaster dokumen itu berada [8].

Untuk mendapatkan nilai keinformatifan suatu *term* pada *cluster* dokumen, maka diperlukan klasterisasi pada dokumen. *Clustering* dokumen-dokumen *corpus* dalam melakukan pembangunan thesaurus secara otomatis dapat membantu meningkatkan relevansi antar *term*, dikarenakan dokumen-dokumen tersebut berkelompok-kelompok berdasarkan karakternya yang sama sehingga *term-term* pada dokumen di *cluster* yang sama juga memiliki nilai kemiripan yang tinggi. Sehingga *term-term* pada *cluster* yang berbeda merupakan *term-term* yang memiliki topik bahasan yang berbeda yang mengakibatkan nilai relevansinya jauh.

Oleh sebab itu paper ini mengusulkan metode baru dalam perhitungan kemiripan *term* dengan *co-occurrence* yang memperhatikan *cluster* dari dokumen pada pengembangan thesaurus Bahasa Arab. Metode yang diusulkan merupakan pengembangan dari metode perhitungan *co-occurrence* sebelumnya, di mana metode sebelumnya memperhatikan kemunculan bersama *term* berdasarkan dokumen [2][3]. Sedangkan metode yang diusulkan memperhatikan kemunculan bersama *term* berdasarkan *cluster* dari dokumen dengan memanfaatkan perhitungan *Inverse class frequency*. *Inverse class frequency* digunakan untuk melihat nilai keinformatifan suatu *term* pada cluster dokumen. Metode yang disusulkan pada paper ini digunakan untuk meningkatkan relevansi *term-term* pada thesaurus.

II. METODE PENELITIAN



Gambar 1. Alur Proses Metode Usulan

Gambar 1 merupakan alur dari metode yang diusulkan. Metode tersebut merupakan pengembangan metode sebelumnya di mana metode sebelumnya memperhatikan kemunculan bersama *term* berdasarkan dokumen dengan memanfaatkan perhitungan TF-IDF, sedangkan metode usulan memperhatikan *cluster* dokumen dengan memanfaatkan perhitungan TF-ICF.

A. *Data*

Dokumen yang digunakan merupakan dokumen-dokemen fiqih berbahasa arab yang diambil dari E-Book pada Maktabah Syamilah. Dokumen yang digunakan sebanyak 1000 dokumen. Di mana dokumen tersebut diambil dari 5 topik yang berbeda-beda yaitu haji, nikah, puasa, sholat dan zakat. Contoh dari dokumen yang digunakan dapat dilihat pada Gambar 2.

ابنته بدون صداق مثلها فقد ثبت النكاح بالمسمى وان فعل ذلك غير الأب ثبت النكاح وكان لها مهر مثلها ومن زوج علاما غير بالغ أو معتوها لم يجز إلا أن يزوجه والده أو وصي داظر له في التزويج وإذا زوج امة بغير إذنها لزمها النكاح وان كرهت كبيره كانت أو صغيره وان زوج عبده وهو كاره لم يجز إلا أن يكون صغيرا وإذا زوج الوليان فالنكاح للأول منهما فإن دخل بها الثاني وهو لا يعلم أنها ذات زوج فرق بينهما وكان لها عليه مهر مثلها ولم يصبها زوجها حتى تحيض ثلاث حيض بعد آخر وقت وطئها الثاني وان جهل الأول منها فسخ النكاحان وإذا تزوج العبد بغير إذن سيده فالنكاح باطل وإن دخل بها فعلى سيده خمسا المهر كما قال عثمان بن عفان رضي الله عنه إلا أن يجاوز الخمسان قيمته فلا يلزم سيده أكثر من قيمته أو يسلمه وإذا تزوج الأمة على إنها حره وأصابها فولدت منه فالولد حر وعليه أن يفديهم والمهر المسمى ويرجع بذلك كله على من غره ويفرق بينهما إن لم يكن له ممن يجوز له أن ينكح الإماء وان ما كان ممن يجوز له أن ينكح فرضي بالمقام فما ولدت بعد الرضي فهو رقيق وان كان المغرور عبدا فولده أحرار ويفديهم إذا عتق ويرجع

Gambar 2. Contoh Dokumen Bahasa Arab

B. *Preprocessing*

Pada tahapan ini dilakukan *preprocessing* dari dokumen untuk mendapatkan *term-term* yang siap diolah pada proses berikutnya. Dokumen-dokumen fiqih berbahasa arab yang telah dikumpulkan akan di proses pemisahan rangkaian kata berdasarkan delimiter atau pemisah kata seperti karakter spasi. Proses pemisahan tersebut sering disebut dengan *tokenizing*. Kemudian proses dilanjutkan dengan *normalization* dan *filtration* yaitu menghilangkan harokat serta simbol-simbol yang tidak penting. Kemudian proses berlanjut dengan *stopword removal* atau menghapus kata-kata yang dianggap tidak penting. Untuk mendapatkan bentuk kata dasar maka kata-kata atau yang disebut dengan term dilakukan *stemming*.

C. *Clustering Dokumen*

Tahapan *clustering* digunakan untuk mengelompokkan dokumen-dokumen berdasarkan kedekatannya. Dokumen-dokumen *corpus* yang banyak dan tidak memiliki label perlu di *clustering* untuk dikumpulkan dengan dokumen-dokumen yang sejenis. Metode untuk melakukan *clustering* dapat dibagi dua yaitu secara hirarki dan secara partisi. Untuk melakukan *clustering* pada dokumen, metode partisi sangat cocok karena kebutuhan komputasi yang rendah. Sedangkan metode hirarki memiliki kompleksitas waktu yang tinggi [9].

Metode *clustering* K-Means merupakan salah satu metode partisi yang sangat mudah untuk diimplementasikan serta memiliki waktu kompleksitas yang rendah. Metode K-Means dilakukan dengan memilih K-dokumen sebagai *centroid*. Kemudian menghitung jarak setiap dokumen terhadap dokumen-dokumen *centroid*. Perubahan titik *centroid* dilakukan secara berulang hingga tercapainya *stop criteria* [9].

Perhitungan jarak dokumen dengan titik *centroid* dilakukan dengan konsep *vector space model* di mana perhitungannya menggunakan *cosine similarity* seperti pada persamaan (1) di mana d merupakan titik *centroid* cluster ke a, sedangkan d' merupakan dokumen ke i [9].

$$\cos(d, d') = \frac{d \cdot d'}{|d||d'|} \quad (1)$$

Perubahan titik *centroid* dilakukan setelah semua dokumen yang ada telah terbagi ke dalam klaster-klaster. Perubahan titik *centroid* akan berhenti hingga *stop criteria* terpenuhi. *Stop criteria* dapat terjadi bila perubahan titik *centroid* tidak signifikan atau telah didefinisikan di awal batas perulangan dari proses clusterisasi itu sendiri. Dalam perubahan titik *centroid* cluster dilakukan dengan mengikuti persamaan (2) di mana d_i merupakan vektor dokumen pada cluster s_j. c_j merupakan vektor *centroid* sedangkan n_j merupakan jumlah dokumen yang terdapat pada klaster s_j [10].

$$c_j = \frac{1}{n_j} \sum_{\forall d_i \in s_j} d_i \quad (2)$$

Clustering pada dokumen-dokumen *corpus* dilakukan dengan *stoping criteria* 1000 iterasi. Dari percobaan yang telah dilakukan dengan berbasis pada *elbow method* didapatkan 6 *cluster* dokumen. *Cluster*

dokumen tersebut yang digunakan untuk proses selanjutnya.

D. Term Frequency – Inverse Class Frequency

Inverse class frequency yang disingkat menjadi ICF merupakan salah satu metode pembobotan *term*. Pembobotan *term* dengan ICF memperhatikan kemunculan *term* pada kumpulan kategori atau *cluster*. *Term* yang jarang muncul pada banyak *cluster* adalah *term* yang bernilai untuk klasifikasi. Kepentingan tiap *term* diasumsikan memiliki proporsi yang berkebalikan dengan jumlah kelas yang mengandung *term* [8].

Pembobotan ICF dapat mengetahui nilai keinformatifan suatu term pada cluster. Hal ini dibuktikan oleh Fauzi dkk serta Septiyawan dkk, di mana ICF digunakan untuk pembobotan term pada perangkingan dokumen [8] [11].

TF-ICF dilakukan dengan melihat frekuensi *term* terhadap *cluster* sesuai dengan persamaan (3) di mana nilai ICF pada *term j* dipengaruhi dengan jumlah *cluster* yang ada *C*, dan cf_j jumlah *cluster* yang mengandung *term j*.

$$ICF_j = 1 + \log \frac{C}{cf_j} \quad (3)$$

Kemudian setiap *term j* pada *cluster-cluster i* akan dihitung bobotnya C_{ji} dengan perhitungan pada persamaan (4) di mana tf_{ji} frekuansi *term j* pada cluster *i* dikali dengan ICF *term j*.

$$C_{ji} = tf_{ji} \times ICF_j \quad (4)$$

E. Co-occurrence Berdasarkan Cluster Dokumen

Setelah dokumen-dokumen fiqih Bahasa Arab di *clustering* menggunakan metode K-Means dan setiap term memiliki nilai bobot TF-ICF, tahapan yang dilakukan selanjutnya adalah melakukan tahapan teknik *co-occurrence* pada *term-term* yang ada. Teknik *co-occurrence* ini dilakukan dengan perhitungan *clusterweight*.

Perhitungan *clusterweight* yang dilakukan pada paper ini merupakan pengembangan dari perhitungan *clusterweight* konvensional [3]. Tujuan dari perhitungan *clusterweight* pada paper ini melakukan pembobotan berdasar kan *cluster* dokumen atau TF-ICF. Perhitungan kemiripan tersebut dilakukan terhadap semua *term* yang ada dengan mengikuti persamaan (5) yang merupakan perhitungan kemiripan term pada tahapan teknik *co-occurrence* berbasis pada ICF di mana C_{hjk} merupakan bobot *term j* dan *term k* muncul bersama pada *cluster h* (TF-ICF *term j* dan *term k)*. TF didapat dari frekuensi terkecil dari kedua *term* tersebut. C_{hj} merupakan bobot *term j* pada *cluster h (*TF-ICF *term j)*. Sedangkan Cf_k merupakan jumlah *cluster* yang terdapat *term k* dan M adalah jumah *cluster* keseluruhan.

$$ClusterWeight(t_j, t_k) = \times \frac{\sum_{h=1}^{m} C_{hjk}}{\sum_{h=1}^{m} c_{hj}} \times \frac{\log \frac{M}{Cf_k}}{\log M} \quad (5)$$

Diketahui pula bahwa perhitungan ini merupakan perhitungan statistikal asimetris di mana $ClusterWeight(t_j, t_k) \neq ClusterWeight(t_k, t_j)$.

Hasil dari perhitungan *clusterweight* merupakan nilai kemiripan antara term *j* dan term *k*. Nilai kemiripan tersebut digunakan untuk membentuk thesaurus.

III. HASIL PENELITIAN

Perhitungan *clusterweight* menghasilkan nilai kemiripan antar *term* di mana nilai tersebut digunakan untuk membentuk thesaurus. Contoh hasil kemiripan *term* yang dijadikan thesaurus jika digambarkan dapat dilihat pada Tabel 1.

Tabel 1. Contoh Hasil Thesaurus

صلي		
Doa		
سلم	ولي	ملك
Salam	Wakil	Raja

Tabel 2. Daftar *Query* Pengujian

ID	QUERY
Q1	متنوعة في وئام في الصلاة " Rukun-rukun dalam sholat"
Q2	كيفية تنفيذ القانون من الزواج " Bagaimana hukum melaksanakan nikah"
Q3	شروط الحج " Syarat-syarat melaksanakan ibadah haji"
Q4	لماذا ينبغي أن أداء الحج " Mengapa harus melaksanakan ibadah haji"
Q5	أمور محرمة في الحج " Hal-hal yang dilarang saat berhaji"
Q6	أي شخص ملزم يصدر الزكاة " Siapa saja yang wajib mengeluarkan zakat"
Q7	المراسيم عشر " Tata cara berzakat"
Q8	كيف يمكن للقانون الصوم " Bagaimana hukum dalam berpuasa"
Q9	ما هي الأشياء التي تفطر الصائم فقط " Hal-hal apa saja yang membatalkan puasa"
Q10	لماذا يجب الصيام " Mengapa harus berpuasa"

Pengujian terhadap hasil pembentukan thesaurus dilakukan dengan menerapkannya terhadap pencarian dokumen. Untuk pengujian ini digunakan *query* yang telah ditentukan seperti pada Tabel 2. *Query* yang telah ditentukan tersebut dimasukkan kedalam sistem untuk mendapatkan hasil perankingan dokumen fiqih bahasa arab.

Term-term pada *query* akan dilihat kemiripannya dengan *term* lain pada thesaurus yang telah dibentuk. *Term-term* yang mirip akan digabungkan dengan *term*

pada *query* untuk digunakan dalam pencarian dokumen. Hasil dari pengujian didapatkan seperti pada Tabel 3.

Tabel 3. Hasil Pengujian

Query	Metode Usulan			*Co-occurrence*		
	P (%)	R (%)	F (%)	P (%)	R (%)	F (%)
Q1	20,0	17,6	18,8	50,0	44,1	46,9
Q2	10,0	27,3	14,6	34,1	63,6	34,1
Q3	50,0	**78,6**	**50,0**	30,0	64,3	40,9
Q4	26,7	53,3	35,6	40,0	80,0	53,3
Q5	3,3	14,3	5,4	10,0	42,9	16,2
Q6	23,3	77,8	35,9	23,3	77,8	35,9
Q7	**63,3**	30,2	40,9	50,0	23,8	32,3
Q8	11,8	9,1	11,8	63,3	34,5	44,7
Q9	23,3	50,0	31,8	20,0	42,9	27,3
Q10	10,0	42,9	16,2	10,0	42,9	16,2

IV. PEMBAHASAN

Dilihat dari tabel hasil pengujian, nilai-nilai *precision, recall* dan *f-measure* hasil pengujian banyak yang bernilai kecil atau di bawah 50%, hal ini dikarenakan oleh beberapa faktor. Yang pertama adalah metode *query expansion* di mana *term-term* pada thesaurus ditambahkan secara langsung terhadap *term-term query* hal tersebut menyebabkan bobot term *query* asli dengan *term* thesaurus menjadi sama dan mengubah nilai informasi dari *query* tersebut.

Faktor kedua adalah perseberan *term* terhadap *cluster* yang ada di mana suatu term memiliki frekuensi yang besar pada setiap *cluster* namun memiliki makna atau konteks yang berbeda mengingat kategori asli pada dokumen yang digunakan merupakan kategori-kategori yang memiliki konteks yang berbeda. Permasalahan tersebut dapat disiasati dengan lebih mengkonvergenkan kategori pada dokumen-dokumen *corpus* dan juga menggunakan metode *query expansion* yang tidak mengurangi nilai keinformatifan *query* awal.

Dilihat pula pada hasil dokumen yang di-*retrieve*. Di mana dokumen-dokumen tersebut memiliki topik yang sama yaitu topik "puasa", hal tersebut dapat diartikan bahwa *term-term* yang dihasilkan pada thesaurus berada pada satu topik. Meskipun beberapa *term* juga terdapat pada topik lain, namun jika *term-term* hasil thesaurus digabungkan sesuai *term query* maka akan menunjuk pada topik yang sama. Dari sini dapat disimpulkan bahwa term-term hasil thesaurus berkelompok dan relevan dengan topik-topik yang ada. Namun perlu adanya filterisasi atau *treshold* terhadap hasil perhitungan *clusterweight* yang tepat untuk melakukan pembentukan thesaurus, karena semakin kecil nilai *treshold* pada hasil *clusterweight* akan menyebabkan banyaknya *term* yang ikut digunakan pada *query expansion*. Sehingga makna dari *query* asli akan berubah.

V. PENUTUP

A. Kesimpulan

Metode pendekatan statistikal asimetris *co-occurrence* berdasarkan *cluster* dokumen merupakan metode pengukuran kemiripan term pada pembentukan thesaurus yang memperhatikan frekuensi kemunculan bersama dilihat dari cluster dokumen. Metode usulan tersebut dapat membentuk thesaurus Bahasa Arab secara otomatis. Hal ini dibuktikan dengan nilai terbesar *precision* sebesar 63,3% di mana nilai tersebut mampu bersaing dengan nilai *precision* dari metode *co-occurence* konvensional.

B. Saran

Dalam pembentukan thesaurus secara otomatis dengan pendekatan statistikal perlu memperhatikan kategori thesaurus yang akan dibentuk. Di mana kategori yang dibentuk harus spesifik seperti contoh pembentukan thesaurus kategori politik dan lain sebagainya. Dengan spesifikasi kategori tersebut maka *term-term* yang terbentuk merupakan *term* yang memiliki konteks yang sama.

DAFTAR PUSTAKA

[1] M. Otair, D. Ph, J. Amman, R. Kanaan, and D. Ph, "Optimizing an Arabic Query using Comprehensive Query Expansion Techniques," *Int. J. Comput. Appl.*, vol. 71, no. 17, pp. 42–49, 2013.

[2] Y. Tseng, "Automatic Thesaurus Generation for Chinese Documents," *J. Am. Soc. Inf. Sci. Technol.*, vol. 53, no. September, pp. 1130–1138, 2002.

[3] H. Khafajeh, M. Refai, and N. Yousef, "Building Arabic Automatic Thesaurus Using Co-occurence Technique," in *Proceedings of International Conference on Communication, Media, Technology and Design*, 2013, pp. 28–32.

[4] P. Li, H. Wang, K. Q. Zhu, Z. Wang, and X. Wu, "Computing term similarity by large probabilistic isA knowledge," *Proc. 22nd ACM Int. Conf. Conf. Inf. Knowl. Manag. - CIKM '13*, pp. 1401–1410, 2013.

[5] E. W. Y. Ismail, "Aplikasi Berbasis Web Pendeteksi *Plagiarisme* Menggunakan Algoritma Himpunan Kata," J. INFOTEL, vol. 6, no. 2, pp. 2–7, 2014.

[6] H. Zohar, C. Liebeskind, J. Schler, and I. D. O. Dagan, "Automatic Thesaurus Construction for Cross Generation Corpus," *J. Comput. Cult. Herit.*, vol. 6, no. 1, 2013.

[7] Y. H. Tseng, "Automatic thesaurus generation for Chinese documents," *J. Am. Soc. Inf. Sci. Technol.*, vol. 53, no. 13, pp. 1130–1138, 2002.

[8] M. A. Fauzi *et al.*, "Term Weighting Berbasis Indeks Buku Dan Kelas Untuk Perangkingan Dokumen Berbahasa Arab," *Lontar Komput.*, vol. 5, no. 2, pp. 110–117, 2015.

[9] M. Mahdavi and H. Abolhassani, "Harmony K -means algorithm for document clustering," no. November 2008, pp. 370–391, 2009.

[10] H. Gupta and R. Srivastava, "k-means Based Document Clustering with Automatic ' k ' Selection

and Cluster Refinement," *Int. J. Comput. Sci. Mob. Appl.*, vol. 2, no. 5, pp. 7–13, 2014.

[11] S. R. Wardhana, D. R. Yunianto, A. Z. Arifin, and D. Purwitasari, "Pembobotan Kata BerbasiS Preferensi Dan Hubungan Semantik Pada Dokumen Fiqih Berbahasa Arab," *J. Teknol. Inf. dan Ilmu Komputer.*, vol. 2, no. 2, pp. 132–137, 2015.